An Automated Process to Create Preservation and Publishing Copies of Digitized Works at the BND

José Borbinha[1], João Gil[2], Gilberto Pedrosa[3], João Penas[4]

INESC-ID – Instituto de Engenharia de Sistemas e Computadores, Rua Alves Redol 9,
Apartado 13069, 1000-029 Lisboa, Portugal
[1]jlb@ist.utl.pt, [2]jgil@ext.bn.pt, [3]gfsp@ext.bn.pt, [4]jpenas@ext.bn.pt

Abstract. This paper describes the automated process to create structured master and access copies for the digitised works at the BND – National Digital Library. The BND created during 2004 and 2005 nearly half a million of digitized images, from more than 25.000 titles of printed works, manuscripts, drawings and maps. The resulting of the digitisation process is a group of TIFF image files representing the surfaces of the original works, which needs yet to be processed in order to be stored and published. Doing that manually would be a very complex and expensive task, with risks for the uniformity of the results, so it was need to develop an automated solution. To create the technical metadata, apply image processing actions and OCR, create derived copies for access in PNG, JPG, GIF, and PDF, we developed a tool named SECO. To create the master copies for each of those works, for preservation, and access copies in HTML, we developed a tool named ContentE, which exists as a standalone tool and as a library. Finally the copies are deposited and registered at the BND repository through the service PURL.PT, which assures also the WEB and intranet access control. This complex process is fully automated through several XML schemas for the control of the processes, description of the results (including the OCR outputs), descriptive metadata (in Dublin Core, MARC XML, etc.) and rights and structural metadata (in METS).

1 Introduction

This paper describes the automated process to create structured master and access copies for the digitised works at the BND – National Digital Library [3]. To create the technical metadata and apply image processing actions, we developed a tool named SECO. To create the master and access, we developed a tool named ContentE. Finally the copies are deposited and registered at the BND repository through the service PURL.PT, which assures also the WEB and intranet access control.

All these components interoperate using XML schemas for the control of the processes and description of the results This is fundamental to meet the BND requirement of an open and widely accepted mean to exchange information with heterogeneous systems. Standardized toolsets and libraries are continuously under development, and XML is the most suitable technology for both data representation and storage. In particular, modern metadata standards and toolsets, such as for METS and Dublin Core, are essentially “XML based”.

The next section describes the overall architecture for the system, stressing the main components. Following, we describe the main relevant sub-systems by their sequence in the workflow (SECO, ContentE and PURL.PT). Finally, we present some conclusions and ideas for future work.

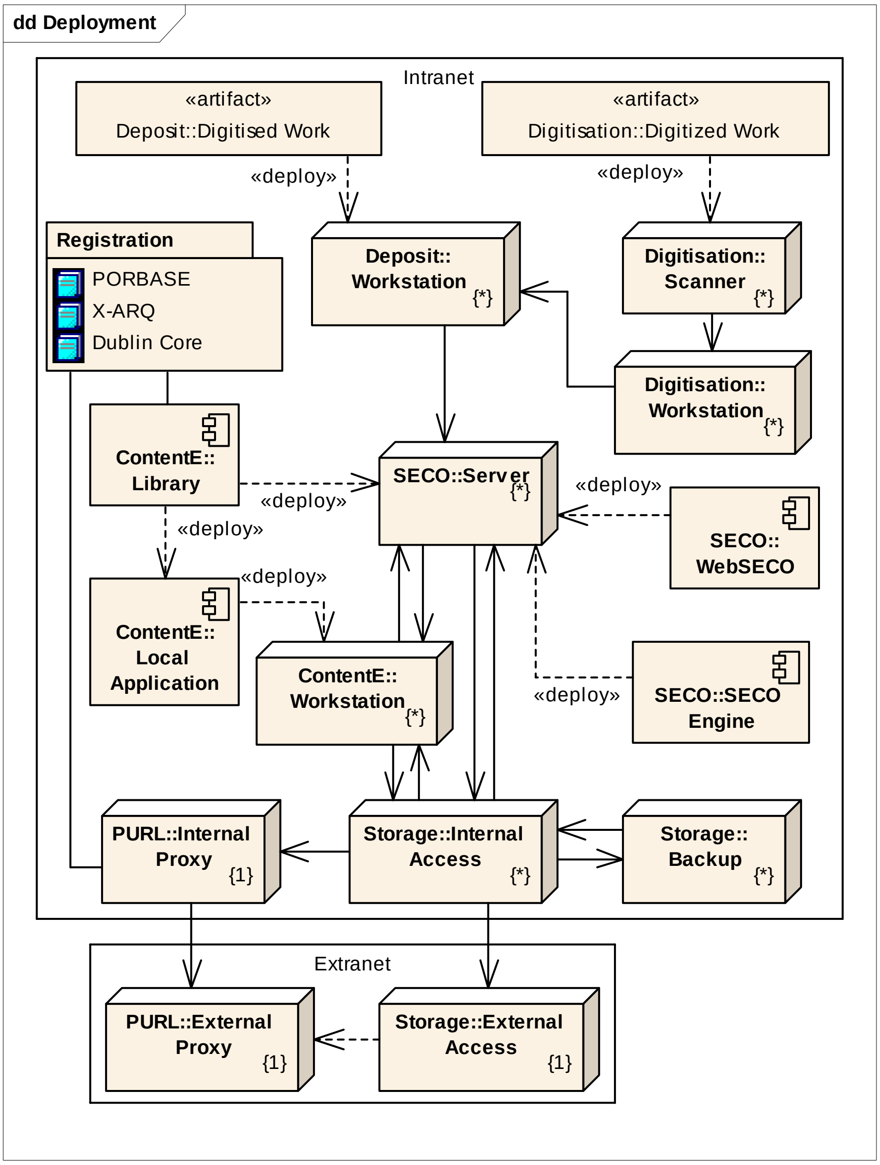


Fig. 1. The overall architecture of the information system supporting the image processing, metadata creation and copies biding for the digitized works at the BND

2 About the overall architecture of the BND

The main purposes of the BND are the development of services for the preservation, registration, discovery and access to digital resources. Those resources comprise digitised and digital born cultural and scientific documental resources. In this paper we are focused only in problem of the digitised resources.

The BND created during 2004 and 2005 nearly half a million of digitized images, from more than 25.000 physical items, comprising printed works, manuscripts, drawings, maps, etc. The resulting of the digitisation of each physical item is a group of TIFF image files representing each surface of the item. In this project nearly 50% of the images, those of items with only text and with good detail, were digitised at the resolution of 300 dpi. The other 50% were digitised at 600 dpi. All the images have 24 bit of colour dept. The biggest images, representing large maps with sizes closer to A1, have more than 2 GBytes.

The master images are perfect for preservation and high quality reproduction, but are too much for access, especially in the Internet. Also, each group of master images representing the same original needs to be described according to that before can be sent to storage. All of this represents a very heavy processing. Doing that manually would be a very complex and expensive task, with risks for the uniformity of the results, so it was identified the need to develop an automated solution. This resulted in a system made of multiple interoperable components, as represented in the Fig. 1 (this image shows only the architecture and components relevant for the scope of this paper, as the overall architecture of the BND comprised also other components here not represented).

In this description we can see that all the process starts with the submission of a digitised work, which can be created internally at the BN, or by an external entity and deposited at the BN. After any of those works is submitted to a local workstation, and after an initial simple inspection, it is sent to SECO. SECO is a service for the creation of the technical metadata, applying of image processing actions, including OCR, as also the creation of derived copies for access in PNG, JPG, GIF, and PDF.

After the processing in this system, SECO calls the ContentE library to create the METS structure and the preservation and access copies. The results are then inspected by a professional, and if they accepted, they are sent directly to storage. However, if it is detect any unusual problem, or if it is decided to publish a specific work in a fancier way, we can use a ContentE Workstation for that. Here the structure can be tuned trough a powerful user interface (creating more indexes and applying a customised style sheet).

Finally the copies are deposited and registered at the BND repository through the service PURL.PT, which assures also the WEB and intranet access control. The PURL.PT service interoperates also with the other traditional registration services, such as the PORBASE – *Base Nacional de Dados Bibliográficos* [9] for the bibliographic works, X-ARQ for the archival works, etc.

This complex process is fully automated through several XML schemas for the control of the processes, description of the results (including the OCR outputs), descriptive metadata (in Dublin Core [10], UNIMARC, coded in MARCXML [7], etc.), as also rights and structural metadata, expressed in METS [8].

3 SECO – Serial Converter

The publishing process of digitized works at BND is essentially geared towards online availability. Non-standard formats and platform dependencies are avoided so that the widest possible range of devices and software can access the information (this is also an important requirement for long term preservation, anyway). For public web-based access, data formats are carefully chosen in order to provide good quality within current technological limitations – bandwidth in particular. On the local intranet (inside the network of the BN – Biblioteca Nacional), however, quality can and should be optimized through larger files and richer content.

To achieve these goals, the publishing process begins with the submission of master high-resolution images to the SECO system. These are usually acquired through top quality digitization techniques and equipment and stored as uncompressed TIFF files at 24 bit colour depth. The typical resolution is 600dpi, or at least 300dpi for good quality printed works containing only text. Consequently, the resulting files can be quite large, often reaching sizes above 1 GByte, which is obviously a major concern in the system.

At its core, the SECO system performs image processing operations. The most basic of these operations is format conversion. TIFF files are not commonly supported by web browsers, which will be the main access tools for the published works. SECO can generate JPEG, GIF, PNG and PDF image files. Only the latter is not natively supported by current-generation browsers. It is nevertheless a nearly ubiquitous multi-page format and viewers such as the Adobe Acrobat Reader [1] are freely available for most platforms.

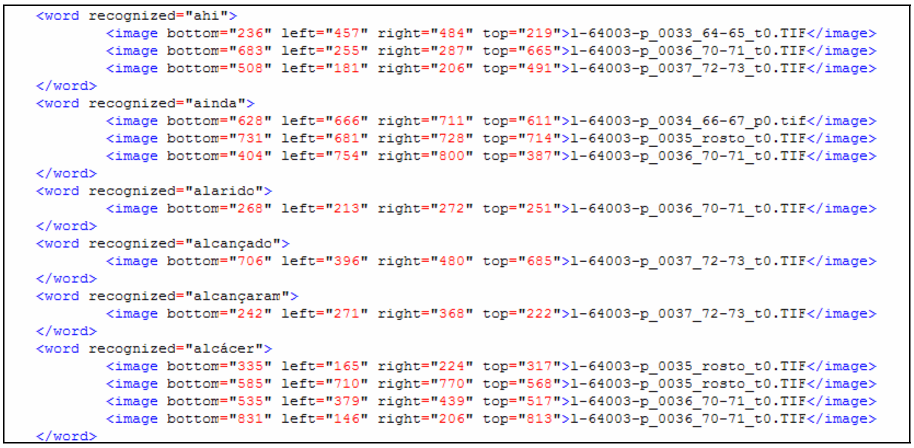
SECO also performs scaling and re-sampling, as well as colour depth reduction. This allows creating manageable image files for consumer-level hardware at 150dpi and 72dpi, common resolutions, and bitmap (1 bit colour depth). These images are also compress very effectively in PNG and the PDF formats. The nature and content of the digitized works must be weighed when selecting these parameters.

Additionally, OCR can be applied to the master images with textual contents, producing standard text files as also textual PDF copies. In general, the results of the OCR are not good enough to be immediately published, and proofreading the extracted text can be very time-consuming (but we do it for some special works). However, this data can be used to build word indexes for the digitized pages. These indexes can later be read by automated indexing tools and complement standard search procedures. Because we store also the word locations (coordinates) in the images, sophisticated presentation mechanisms can be built upon this information. These word indices are kept in XML files, such as represented in the Fig. 2.

For very large and highly detailed images (e.g. maps), creative solutions must be sought to cater for ordinary network connections. For this SECO supports also automated image slicing, this can break a large image into a tree of lightweight parts at different detail levels. This concept is realized afterwards when generating the actual website for the digitised work.

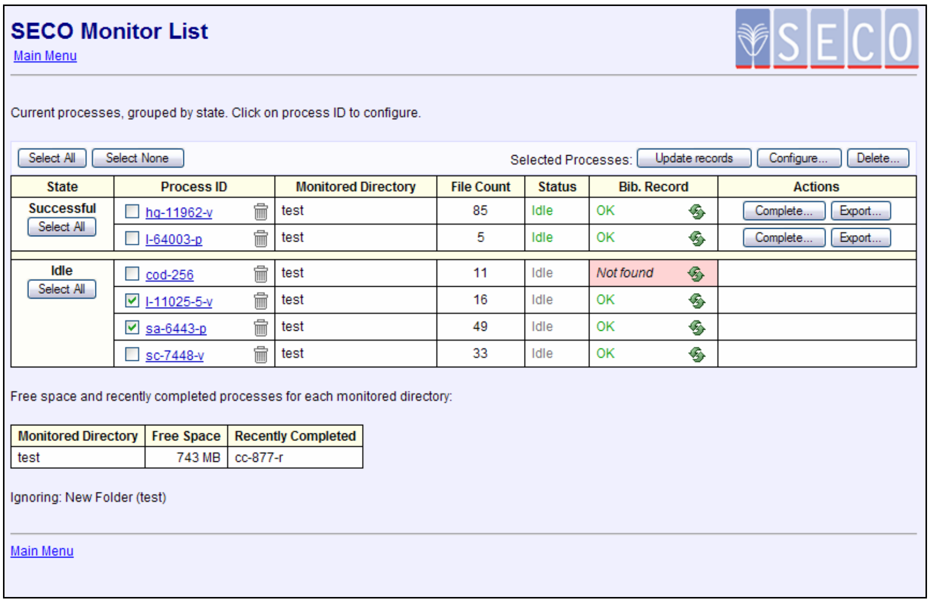
After master images are sent to the SECO server through a network file sharing protocol, all further operations are configured and executed through the SECO web-based user interface known as WebSECO. Fig. 3 shows a screenshot of this interface, representing the process's list with an assortment of available options.

This web interface allows viewing bibliographic records, reports processing status and provides file and process management operations. It also includes a cropping tool to remove unwanted elements from images, such as colour charts and excess margins. The user specifies the crop area graphically on low resolution versions and the system scales the coordinates to the full resolution masters. This way, no manual processing of the heavy high resolution images is necessary. The low resolution copies for this purpose are generated by an optimized rescaling implementation.



```xml
<word recognized="ahi">
        <image bottom="236" left="457" right="484" top="219">1-64003-p_0033_64-65_t0.TIF</image>
        <image bottom="683" left="255" right="287" top="665">1-64003-p_0036_70-71_t0.TIF</image>
        <image bottom="508" left="181" right="206" top="491">1-64003-p_0037_72-73_t0.TIF</image>
</word>
<word recognized="ainda">
        <image bottom="628" left="666" right="711" top="611">1-64003-p_0034_66-67_p0.tif</image>
        <image bottom="731" left="681" right="728" top="714">1-64003-p_0035_rosto_t0.TIF</image>
        <image bottom="404" left="754" right="800" top="387">1-64003-p_0036_70-71_t0.TIF</image>
</word>
<word recognized="alarido">
        <image bottom="268" left="213" right="272" top="251">1-64003-p_0036_70-71_t0.TIF</image>
</word>
<word recognized="alcançado">
        <image bottom="706" left="396" right="480" top="685">1-64003-p_0037_72-73_t0.TIF</image>
</word>
<word recognized="alcançaram">
        <image bottom="242" left="271" right="368" top="222">1-64003-p_0037_72-73_t0.TIF</image>
</word>
<word recognized="alcácer">
        <image bottom="335" left="165" right="224" top="317">1-64003-p_0035_rosto_t0.TIF</image>
        <image bottom="585" left="710" right="770" top="568">1-64003-p_0035_rosto_t0.TIF</image>
        <image bottom="535" left="379" right="439" top="517">1-64003-p_0036_70-71_t0.TIF</image>
        <image bottom="831" left="146" right="206" top="813">1-64003-p_0036_70-71_t0.TIF</image>
</word>
```

Fig. 2. Example of a word index formated in XML



SECO Monitor List

Current processes, grouped by state. Click on process ID to configure.

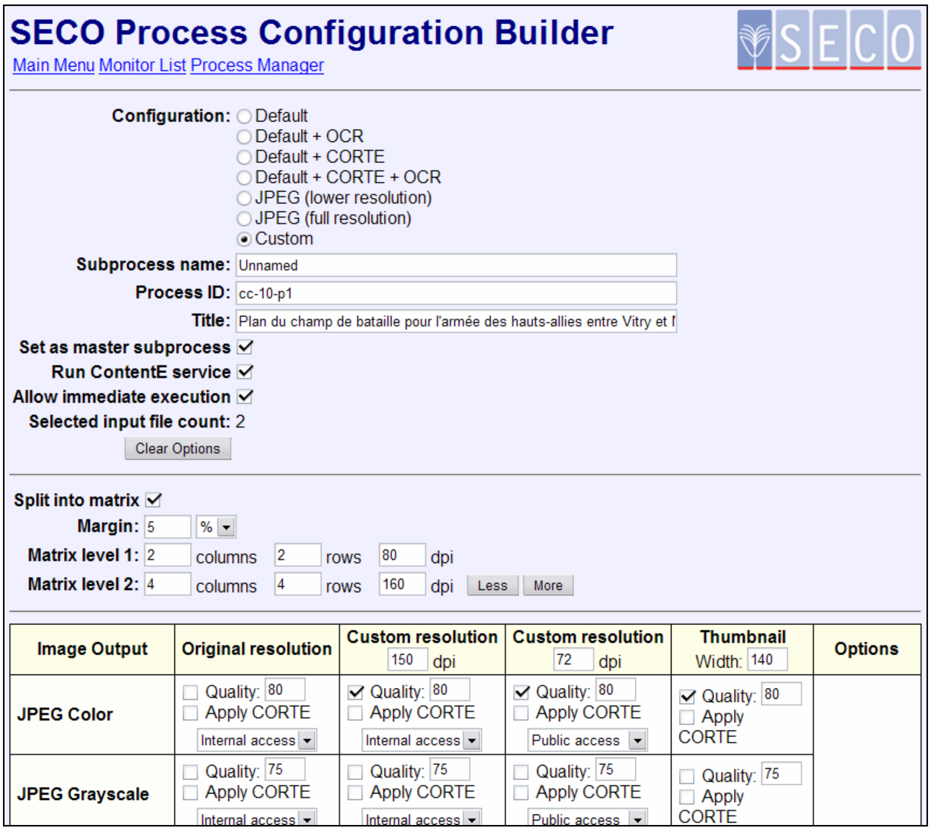
State	Process ID	Monitored Directory	File Count	Status	Bib. Record	Actions
Successful	□ hg-11962-v	test	85	Idle	OK	Complete... Export...
	□ l-64003-p	test	5	Idle	OK	Complete... Export...
Idle	□ cod-256	test	11	Idle	*Not found*	
	☑ l-11025-5-v	test	16	Idle	OK	
	☑ sa-6443-p	test	49	Idle	OK	
	□ sc-7448-v	test	33	Idle	OK	

Free space and recently completed processes for each monitored directory:

Monitored Directory	Free Space	Recently Completed
test	743 MB	cc-877-r

Ignoring: New Folder (test)

Fig. 3. Example of WebSECO, the on-line interface for SECO



SECO Process Configuration Builder

Configuration: ○ Default
○ Default + OCR
○ Default + CORTE
○ Default + CORTE + OCR
○ JPEG (lower resolution)
○ JPEG (full resolution)
● Custom

Subprocess name: Unnamed

Process ID: cc-10-p1

Title: Plan du champ de bataille pour l'armée des hauts-allies entre Vitry et l

Set as master subprocess ☑

Run ContentE service ☑

Allow immediate execution ☑

Selected input file count: 2

Split into matrix ☑

Margin: 5 %

Matrix level 1: 2 columns 2 rows 80 dpi

Matrix level 2: 4 columns 4 rows 160 dpi

Image Output	Original resolution	Custom resolution 150 dpi	Custom resolution 72 dpi	Thumbnail Width: 140	Options
JPEG Color	□ Quality: 80 □ Apply CORTE Internal access	☑ Quality: 80 □ Apply CORTE Internal access	☑ Quality: 80 □ Apply CORTE Public access	☑ Quality: 80 □ Apply CORTE	
JPEG Grayscale	□ Quality: 75 □ Apply CORTE Internal access	□ Quality: 75 □ Apply CORTE Internal access	□ Quality: 75 □ Apply CORTE Public access	□ Quality: 75 □ Apply CORTE	

Fig. 4. The process configuration page of the WebSECO interface

The most complex area of the user interface is, however, the process configuration page, partially shown in Fig. 4. All the available formats and sub-formats are listed in a table, along with compression quality, resolution, slicing and multi-page parameters. Several presets are available as configuration templates for most common cases, usually requiring little or no customization for each individual process. To further enhance productivity, a single configuration can be shared by a group of works, so that a user can quickly setup and launch an arbitrarily large set of similar tasks.

After the configuration is set, the scheduling component can execute the process immediately of only in a pre-determinate moment, avoiding overloading the server with simultaneous processor, memory and disk intensive image operations.

After the processing of the images, the process continues by generating XHTML bindings for the access copies. For this, SECO invokes the ContentE Service. It is an integrated software module on which the ContentE Local Application is also based. A detailed description will be given in the next section.

The final stage at the SECO processing is quality control. If the automated execution results are deemed satisfactory, which is true for the majority of the cases, the finished work is flagged as complete and the SECO system notifies the central storage service, which moves the data to the preservation and online access areas. On the other hand, if the results have insufficient quality, the user can reconfigure the process attempting to correct detected problems. A third option is exporting the generated data to a workspace where it can be manually adjusted, retouched or corrected. This option supports the creation of detailed indexes through the ContentE Local Application.

To store all the data behind the SECO execution, no traditional data base is used. Instead, the system relies on XML extensively, writing information in XML files. This allows for dynamic and flexible data formats and the ability to keep work metadata linked to its main files. Works to be published can be moved, copied and distributed while including relevant properties for indexing, archiving, technical description and additional processing.

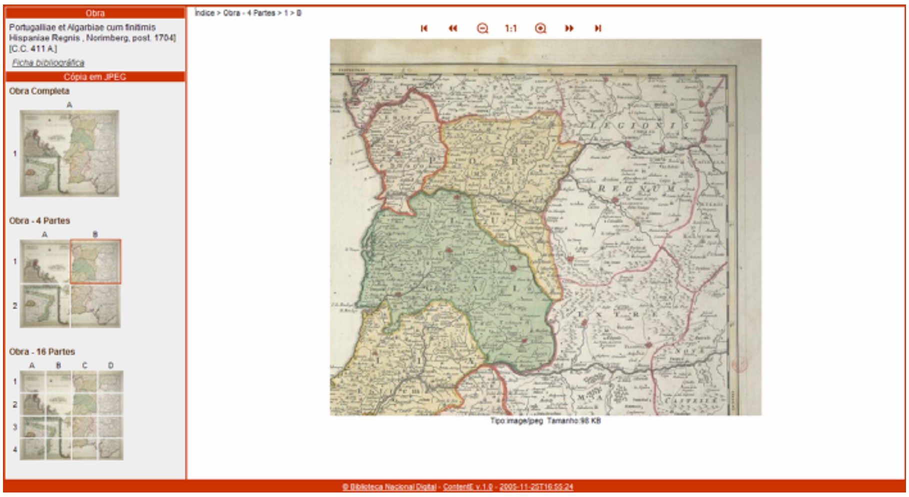
The configurations of the SECO processing core are also stored in XML files. The available presets, parameters, processing modules and configuration templates are themselves described in that way. Than SECO architecture can work as building blocks to fit the needs of each image set. Other information stored by SECO in XML files include pre-processing and crop data, generated file properties and statistics, OCR results, configuration history, process persistence data structures and detailed logs of all the processes.

4 ContentE

The processes based on ContentE are described with detail in [4]. Here we'll make only a resume of it! ContentE is made essentially of two components: a library, to be used by other systems, and a local application that uses the same library and provides a powerful user interface for advanced usage.

The ContentE library is used by SECO to process the results of the previous steps and produce master copies for preservation and copies for access. The master copies are just a folder organised inside in a set of other folders, one for each MIME type existing for the object. One typical MIME type that is always present is TIFF. Other types are usually JPEG, PNG, GIF, PDF and TXT. For all of this, ContentE creates also structural descriptions in METS, as also indexes. One index that is always created automatically is the original physical index, representing the images by their natural order. More complex indexes that can be also automatically created by SECO are tree indexes for images that are split in multiple areas, for better visualisation. Other complex indexes, such as authors, parts, chapters, etc., can be created only using the ContentE Local Application, or they can be also created trough SECO but in this case the respective XML descriptions (in METS) have to be provided with the submission of the images.

ContentE creates one access copy for each derived MIME type. Each of these copies is represented as an XHTML site, where the homepage shows the descriptive metadata and possibly also the indexes and thumbnails of the cover pages, etc. The indexes can be also textual or they can use thumbnails to represent the access points. To create these XHTML copies, ContentE can apply multiple XSL visual styles that are already integrated in the library. Anyway, any new style can be used at any moment, providing that it can be applied to METS. Finally, the ContentE Library can retrieve descriptive metadata from external systems, such as PORBASE and X-ARQ, or import it from local files (in Dublin Core [10] or UNIMARC [11], coded in MARCXML [7]). Each access copy has also an access rights metadata structure, whose parameters can be chosen in the SECO's interface.



Obra

Portugalliae et Algarbiae cum finitimis Hispaniae Regnis , Norimberg, post. 1704] [C.C. 411 A.]

Ficha bibliográfica

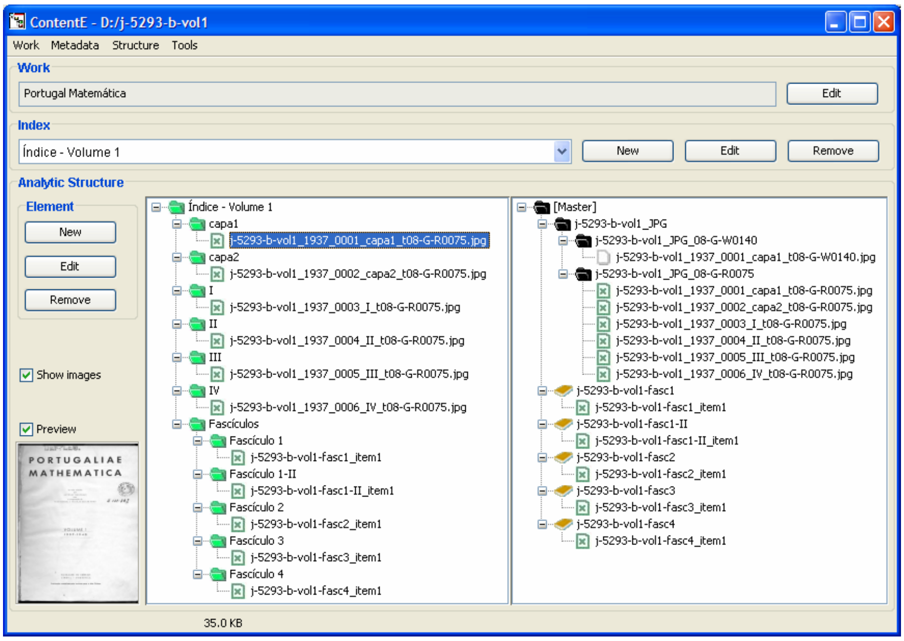
Cópia em JPEG

Índice > Obra - 4 Partes > 1 > B

Tipo:image/jpeg Tamanho:98 KB

© Biblioteca Nacional Digital - ContentE v.1.0 - 2005-11-25T16:55:24

Fig. 5. Example of an access copy of a digitised map structured by ContentE



ContentE - D:/j-5293-b-vol1

Work: Portugal Matemática

Index: Índice - Volume 1

Analytic Structure

☑ Show images

☑ Preview

```
Índice - Volume 1
    capa1
        j-5293-b-vol1_1937_0001_capa1_t08-G-R0075.jpg
    capa2
        j-5293-b-vol1_1937_0002_capa2_t08-G-R0075.jpg
    I
        j-5293-b-vol1_1937_0003_I_t08-G-R0075.jpg
    II
        j-5293-b-vol1_1937_0004_II_t08-G-R0075.jpg
    III
        j-5293-b-vol1_1937_0005_III_t08-G-R0075.jpg
    IV
        j-5293-b-vol1_1937_0006_IV_t08-G-R0075.jpg
    Fascículos
        Fascículo 1
            j-5293-b-vol1-fasc1_item1
        Fascículo 1-II
            j-5293-b-vol1-fasc1-II_item1
        Fascículo 2
            j-5293-b-vol1-fasc2_item1
        Fascículo 3
            j-5293-b-vol1-fasc3_item1
        Fascículo 4
            j-5293-b-vol1-fasc4_item1
```

```
[Master]
    j-5293-b-vol1_JPG
        j-5293-b-vol1_JPG_08-G-W0140
            j-5293-b-vol1_1937_0001_capa1_t08-G-W0140.jpg
        j-5293-b-vol1_JPG_08-G-R0075
            j-5293-b-vol1_1937_0001_capa1_t08-G-R0075.jpg
            j-5293-b-vol1_1937_0002_capa2_t08-G-R0075.jpg
            j-5293-b-vol1_1937_0003_I_t08-G-R0075.jpg
            j-5293-b-vol1_1937_0004_II_t08-G-R0075.jpg
            j-5293-b-vol1_1937_0005_III_t08-G-R0075.jpg
            j-5293-b-vol1_1937_0006_IV_t08-G-R0075.jpg
    j-5293-b-vol1-fasc1
        j-5293-b-vol1-fasc1_item1
    j-5293-b-vol1-fasc1-II
        j-5293-b-vol1-fasc1-II_item1
    j-5293-b-vol1-fasc2
        j-5293-b-vol1-fasc2_item1
    j-5293-b-vol1-fasc3
        j-5293-b-vol1-fasc3_item1
    j-5293-b-vol1-fasc4
        j-5293-b-vol1-fasc4_item1
```

35.0 KB

Fig. 6. Example of a session of the ContentE Local Application structuring a digitised journal.

Fig. 5 shows an example of an access copy of a work produced by SECO and ContentE using these procedures. It is a classical example of a map, for which it as asked to be created an index tree of three levels, where the first shows all the map, the second splits it in four parts (all with the same resolution of the image of the upper level), and the third splits it in 16 images. All was done automatically, once configured in SECO. Fig. 6 shows another example of a work being processed using the ContentE Local Application. In this case it is a journal. In cases like this, we want usually detailed indexes. As SECO produces only sequential indexes, manual work is required for the more complex structuring.

5 PURL.PT

The PURL.PT service allows digital or digitalized works to be registered and accessed through unique and persistent URLs. These URL have the syntax “http://purl.pt/xpto”, where “xpto” is a simple sequential number that started in 1 and grows up for each new work. These identifiers give access to a “home page” of the work, where a user can see descriptive and technical metadata (with references to the MIME formats of each copy), and have also access to the copies. The URLs for the copies have the syntax “http://purl.pt/xpto/copy”, where “copy” is again a sequential number. The value zero is always reserved for the master copy, while the other copies are numbered sequentially without any particular order.

The system has an administration interface, a set of web services to interact with other applications. It works also as a proxy system to allow access control, with the architecture show in Fig. 7.

The works are registered in the PURL.PT service by the submission of a XML file created by ContentE. This file has a METS structure and contains all needed information for register the work and its copies. This file doesn’t contain descriptive information about the copies; it only has references for them. Therefore, each copy has its own METS file with the description of its structure.

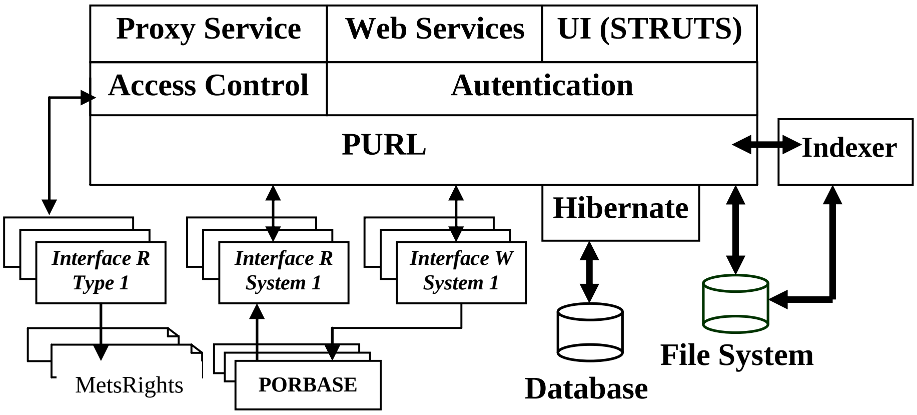
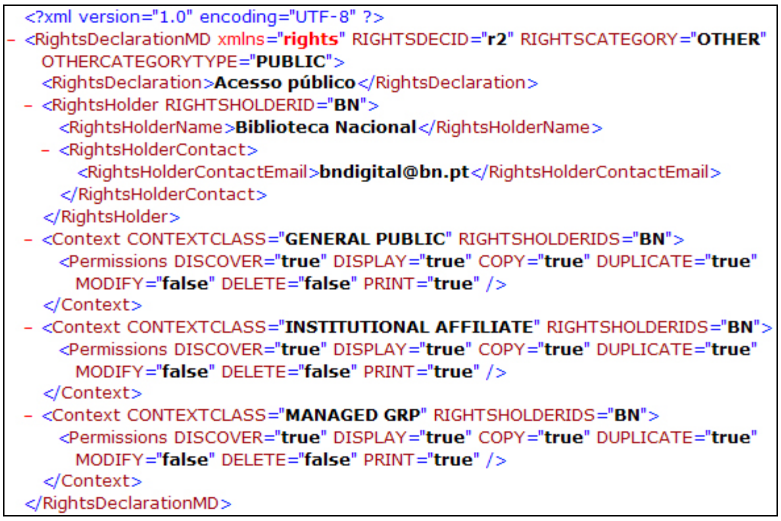
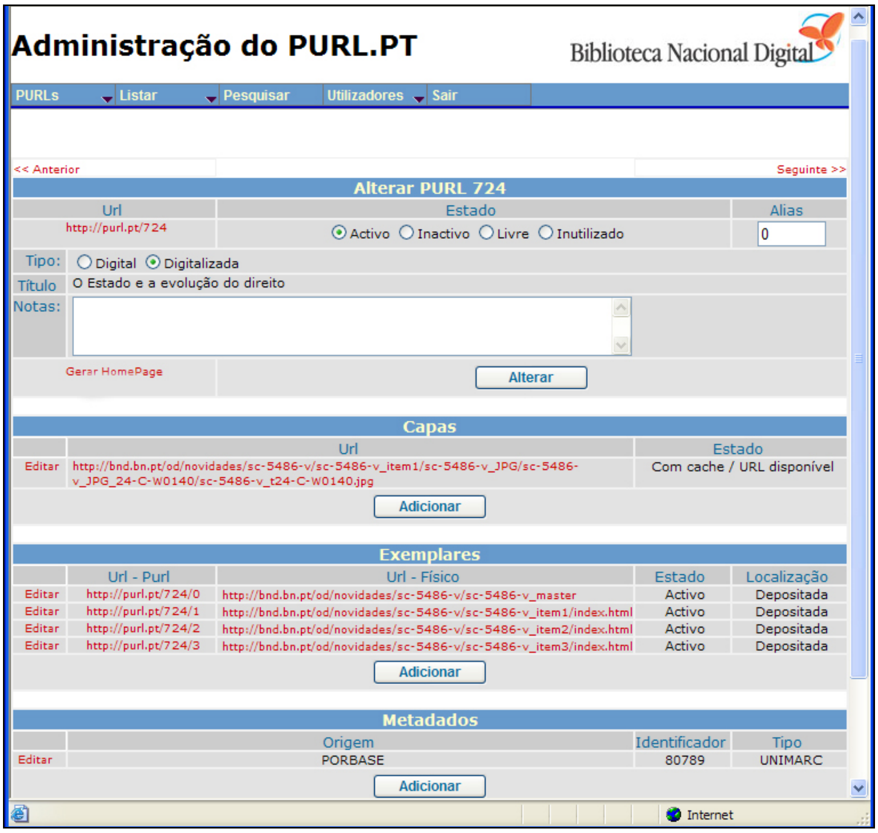


Fig. 7. The architecture of the service PURL.PT



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <RightsDeclarationMD xmlns="rights" RIGHTSDECID="r2" RIGHTSCATEGORY="OTHER"
    OTHERCATEGORYTYPE="PUBLIC">
    <RightsDeclaration>Acesso público</RightsDeclaration>
  - <RightsHolder RIGHTSHOLDERID="BN">
      <RightsHolderName>Biblioteca Nacional</RightsHolderName>
    - <RightsHolderContact>
        <RightsHolderContactEmail>bndigital@bn.pt</RightsHolderContactEmail>
      </RightsHolderContact>
    </RightsHolder>
  - <Context CONTEXTCLASS="GENERAL PUBLIC" RIGHTSHOLDERIDS="BN">
      <Permissions DISCOVER="true" DISPLAY="true" COPY="true" DUPLICATE="true"
        MODIFY="false" DELETE="false" PRINT="true" />
    </Context>
  - <Context CONTEXTCLASS="INSTITUTIONAL AFFILIATE" RIGHTSHOLDERIDS="BN">
      <Permissions DISCOVER="true" DISPLAY="true" COPY="true" DUPLICATE="true"
        MODIFY="false" DELETE="false" PRINT="true" />
    </Context>
  - <Context CONTEXTCLASS="MANAGED GRP" RIGHTSHOLDERIDS="BN">
      <Permissions DISCOVER="true" DISPLAY="true" COPY="true" DUPLICATE="true"
        MODIFY="false" DELETE="false" PRINT="true" />
    </Context>
  </RightsDeclarationMD>
```

Fig. 8. An example of a rights metadata file



Alterar PURL 724

Url	Estado	Alias
http://purl.pt/724	● Activo ○ Inactivo ○ Livre ○ Inutilizado	0

Tipo: ○ Digital ● Digitalizada

Título O Estado e a evolução do direito

Notas:

Gerar HomePage

Capas

	Url	Estado
Editar	http://bnd.bn.pt/od/novidades/sc-5486-v/sc-5486-v_item1/sc-5486-v_JPG/sc-5486-v_JPG_24-C-W0140/sc-5486-v_t24-C-W0140.jpg	Com cache / URL disponível

Exemplares

	Url - Purl	Url - Físico	Estado	Localização
Editar	http://purl.pt/724/0	http://bnd.bn.pt/od/novidades/sc-5486-v/sc-5486-v_master	Activo	Depositada
Editar	http://purl.pt/724/1	http://bnd.bn.pt/od/novidades/sc-5486-v/sc-5486-v_item1/index.html	Activo	Depositada
Editar	http://purl.pt/724/2	http://bnd.bn.pt/od/novidades/sc-5486-v/sc-5486-v_item2/index.html	Activo	Depositada
Editar	http://purl.pt/724/3	http://bnd.bn.pt/od/novidades/sc-5486-v/sc-5486-v_item3/index.html	Activo	Depositada

Metadados

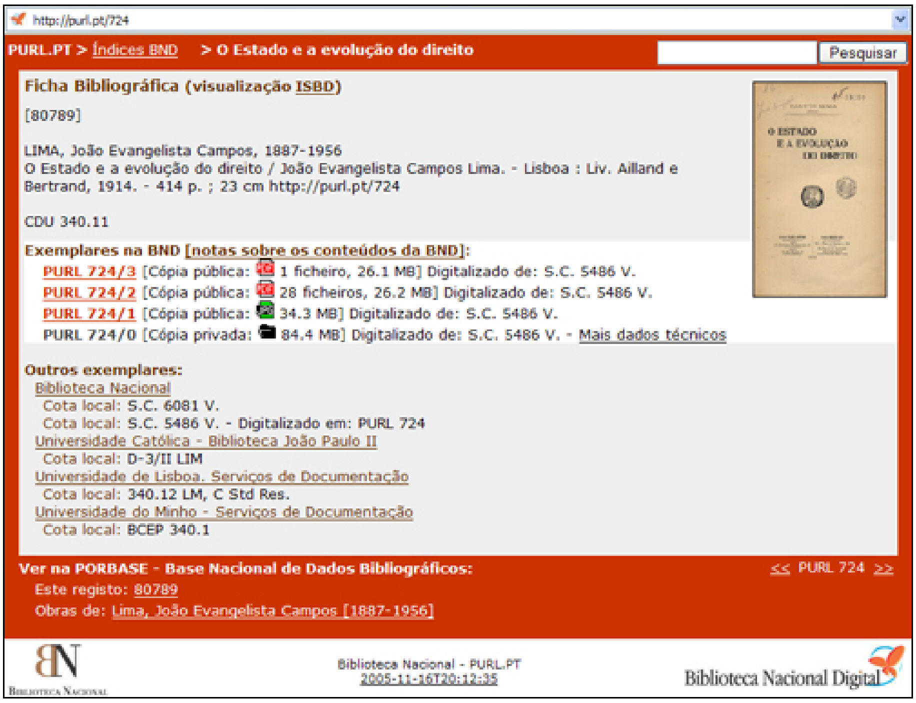
	Origem	Identificador	Tipo
Editar	PORBASE	80789	UNIMARC

Fig. 9. Administrative interface for the service PURL.PT

The PURL.PT service can process multiple descriptive metadata formats (UNIMARC, Dublin Core and others custom). Through the use of different style sheets (according to the format of the record), a descriptive text is created for the homepage of the work. This page also contains information regarding to the work's several digital items and its properties, as well as for its physical items references. The METS' master copy also enables the creation of a HTML page with technical information about the different MIME formats that can be found in the copies.

Each copy contains a XML rights file, referred through the METS file, from which the PURL.PT extracts the terms for access control. In BND we defined tree types of conditions: private, internal or public. Private items can not be accessed by normal users, which are the case of all the master items; internal copies can only be accessed at the intranet; public copies are accessible to all users, including the Internet. The actual schema for this rights format is available online[1], with a sample shown in the Fig. 8.

In the Fig. 9 we can see an example of the administrative interface of the PURL.PT service, and in Fig. 10 we can see an example of a descriptive page created automatically for a digitised work with one master and three copies for access. It is interesting to see also that due to the interoperability with PORBASE, we can see not only the references to the physical item that was originally digitised, but also the references to other copies existing in other libraries members of PORBASE.



PURL.PT > Índices BND > O Estado e a evolução do direito

Ficha Bibliográfica (visualização ISBD)

[80789]

LIMA, João Evangelista Campos, 1887-1956
O Estado e a evolução do direito / João Evangelista Campos Lima. - Lisboa : Liv. Ailland e Bertrand, 1914. - 414 p. ; 23 cm http://purl.pt/724

CDU 340.11

Exemplares na BND [notas sobre os conteúdos da BND]:
PURL 724/3 [Cópia pública: 1 ficheiro, 26.1 MB] Digitalizado de: S.C. 5486 V.
PURL 724/2 [Cópia pública: 28 ficheiros, 26.2 MB] Digitalizado de: S.C. 5486 V.
PURL 724/1 [Cópia pública: 34.3 MB] Digitalizado de: S.C. 5486 V.
PURL 724/0 [Cópia privada: 84.4 MB] Digitalizado de: S.C. 5486 V. - Mais dados técnicos

Outros exemplares:
Biblioteca Nacional
Cota local: S.C. 6081 V.
Cota local: S.C. 5486 V. - Digitalizado em: PURL 724
Universidade Católica - Biblioteca João Paulo II
Cota local: D-3/II LIM
Universidade de Lisboa. Serviços de Documentação
Cota local: 340.12 LM, C Std Res.
Universidade do Minho - Serviços de Documentação
Cota local: BCEP 340.1

Ver na PORBASE - Base Nacional de Dados Bibliográficos:
Este registo: 80789
Obras de: Lima, João Evangelista Campos [1887-1956]

<< PURL 724 >>

Biblioteca Nacional - PURL.PT
2005-11-16T20:12:35

Fig. 10. Descriptive page for a digitised work created automatically by the PURL.PT service

[1] http://schemas.bn.pt/right/v1/rightsv1.xsd

6 Conclusions

In this paper we explained how we harnessed a very complex problem with a strategy based on the development of a processing system integrating several functional blocks by the effective use of XML in the interfaces and transport of data. The overall system is actually in production at the BND, and the experiences had so far make us to believe that it'll make it possible to publish, in a short time and with fair human intervention, the more than 20.000 digitised titles that are still waiting in the queue.

All the code was developed in JAVA, and all the results are available in open-source. We have also plans to continue the developments, especially to reinforce SECO with more image processing features (such as to cut margins and improve the legibility in images), add more schemas to ContentE (such as the DOCBOOK [6] and the Digital Talking Book [2] defined by the DAISY Consortium [5], which will make it possible to process also object with sound, for visual impaired persons). Finally, the PURL.PT service has many other features that were not mentioned in this paper, especially to support browsing in the BND and the publishing of collections and profiles, all supported easily thanks to the global usage of well defined XML schemas. All of this will make it possible to support new services, such as, for example, a user space under development, where users will be able to register and take advantage of new customised services.

References

[1] Adobe Acrobat Reader <http://www.adobe.com/products/acrobat/>

[2] ANSI/NISO Z39.86-2005. Specifications for the Digital Talking Book. ISSN: 1041-5653 <http://www.daisy.org/z3986/2005/z3986-2005.html>

[3] BND. Biblioteca Nacional Digital. <http://bnd.bn.pt>

[4] Borbinha, José; Pedrosa, Gilberto; Penas, João; Gil, João. A gestão de obras digitalizadas na BND. XML: Aplicações e Tecnologias Associadas. 10 e 11 de Fevereiro de 2005, Casa da Torre, Vila Verde, Braga.

[5] DAISY Consortium <http://www.daisy.org/>

[6] DOCBOOK. <http://www.docbook.org/>

[7] MARCXML. MARC 21 XML Schema. <http://www.loc.gov/standards/marcxml/>

[8] METS. Metadada Encoding and Transmission Standard. <http://www.loc.gov/standards/mets>

[9] PORBASE. Base Nacional de Dados Bibliográficos. <http://www.porbase.org>

[10] The Dublin Core Metadata Initiative <http://www.dublincore.org>

[11] UNIMARC <http://www.unimarc.info>